Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis

Shou-de Lin
Computer Science Department
University of Southern California
sdlin@isi.edu

Hans Chalupsky
Information Sciences Institute
University of Southern California
hans@isi.edu

Abstract

A significant portion of knowledge discovery and data mining research focuses on finding patterns of interest in data. Once a pattern is found, it can be used to recognize satisfying instances. The new area of link discovery requires a complementary approach, since patterns of interest might not yet be known or might have too few examples to be learnable. This paper presents an unsupervised link discovery method aimed at discovering unusual, interestingly linked entities in multi-relational datasets. Various notions of rarity are introduced to measure the "interestingness" of sets of paths and entities. These measurements have been implemented and applied to a real-world bibliographic dataset where they give very promising results.

1. Introduction

Link discovery is a relatively new form of data mining with the goal of automatically identifying abnormal or threatening activities in large and heterogeneous data sets. Mooney et al. [10] describe it as the task of "identifying known, complex, multi-relational patterns that indicate potentially threatening activities in large amounts of relational data." Under this view of link discovery, once a pattern of interest is known or has been learned, a sophisticated pattern matcher can use it to detect satisfying instances in the data. The match process is usually difficult given the scale, heterogeneity, distribution, incompleteness and corruption of the data. Its biggest limitation is, however, that it can only detect instances of known patterns and cannot cope with previously unknown or evolving patterns of interest. Senator [17] describes link discovery more broadly as the process of looking for "evidence of known patterns and, perhaps more important, for unexplainable connections that may indicate previously unknown but significant connections, representing, for example, a new group, threat, or capability." It is this requirement for being able to discover novel, previously unknown kinds of links that motivated the work presented in this paper. We will call this requirement the *novel link discovery* (NLD) problem to distinguish it from the overall or more traditional pattern-based link discovery problem (LD).

In the following we describe an unsupervised link discovery approach based on rarity analysis to address the NLD problem. Unsupervised link discovery is different from traditional link discovery from an input/output perspective. A traditional LD program takes multi-relational evidence data and a set of learned patterns as inputs and produces (usually partial) instantiations of the patterns as results. For example, given some police evidence database and a pattern description of contract murders, the program will try to detect and report instances of such murder events. An unsupervised link discovery program takes the same evidence data as input but does not use any pattern information. Instead of pattern instantiations, the results are any interesting connections found in the evidence data based on some model of "interestingness". For example, given the same evidence database the result might be a list of interesting connections between certain criminals or gangs.

Traditionally, knowledge discovery and data mining research focuses on discovering and extracting previously unknown, valid, novel, potentially useful and understandable patterns from lower-level data [20]. Such patterns can be represented as association rules, classification rules, clusters, sequential patterns, time series, contingency tables, etc [9]. Identifying "interesting" information in large, multi-relational data sets without using a pattern, on the other hand, has not received much attention at all. We argue, however, that patterns and rules are not the only things that should be mined from data sets, and that some version of unsupervised, pattern-free link discovery is necessary to handle the NLD problem.

The next section describes the problem and underlying assumptions in more detail. Section 3 defines different rarity measures and how they can be applied to NLD problems. Section 4 describes experiments performed to validate the proposed rarity measures, Section 5 describes related work and in the last section we conclude with a discussion and future work.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Information Sciences Institute ,4676 Admiralty Way,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

2. The problem

In this paper we focus on discovering "interesting" paths and nodes from data that can be represented as sets of entities connected by a set of binary relations. In other words, each object in the data set is treated as a separate entity and there are different types of binary relations connecting these entities. This kind of data can naturally be represented by a labeled graph such as the one shown in Figure 1 where nodes stand for entities and links for binary relations. For example, social network data [16] or Web-pages with proper classification on hyperlinks can be represented in this way. We also assume that the data employs a rich vocabulary of relations where different link types represent different semantic relationships. For example, we have different links representing that "X wrote a letter to Y" or that "X is the brother of Y". Therefore, different graphs with identical structure will usually have very different meanings depending on the types of links involved.

Given these assumptions, we define the following three classes of NLD problems addressed by our approach:

(1) *Novel path discovery:* given an arbitrary pair of entities in a graph, find the most interesting or novel paths between them.

(2) *Novel loop discovery:* given an arbitrary entity in a graph, find the most interesting or novel loops starting and ending at it.

(3) *Significant node discovery:* given an arbitrary entity in a graph, find other entities that are most significantly connected to it. For example, given some person A, find the set of people that A is most significantly connected to.

2.1 Challenges in novel link discovery

The first challenge of the NLD problem is that the term "novelty" or "interestingness" is user dependent. Each person might view data from different angles and, thus, which connections interest them varies. A good NLD program should therefore take users preferences into consideration while still doing most of the work automatically. Balancing this trade-off is a challenging design issue.

The second challenge is that "interestingness" is domain specific, that is, it depends on the characteristics of the particular domain described by the data. For example, for the novel path discovery problem most people would probably think that the link "A killed B" is more interesting than "A wrote a letter to B". The justification for this is that, empirically, the event "killing" happens less frequently than the event "writing a letter". This, however, is only true if the mined data set describes the behavior of the general population. If instead we were looking at a police murder database containing primarily murder events, the reverse would be the case. This is because in this data set everybody is more or less involved in some "killing" event while "writing a letter" is considerably more rare or unusual. In other words, when investigating this data set, users will expect to find data related to "killing" but not necessarily to "writing a letter". Information-theoretically, we can say that "killing" conveys less information in this context. Therefore, the evidence "writing a letter" might surprise a user and trigger him or her to consult additional sources for further information. This explains why it is not sufficient to tackle the NLD problem simply by analyzing individual semantics of the relations, but that it is very important to consider the domain and context the data is in.

A typical supervised learning approach for this problem would be to learn a weight of interestingness for each relation or series of relations in a data set and then apply a shortest path algorithm accumulating these weights to look for solutions. This, however, is not practical due to the difficulty of generating unbiased training data. Take the novel path discovery problem for example. To obtain unbiased training data we have to rank the novelty of training paths manually with consistency. In other words, we have to develop a "standard operating procedure" about how to quantify interestingness of paths in a specific domain. The third challenge arises from this "chicken and egg" dilemma: if we could develop a standard evaluation criterion to judge the interestingness of paths or nodes, then we could apply it directly as our novel link finder and would not have to learn it. But since we do not have such a criterion, we also cannot generate labeled training examples to learn it. This limits the applicability of a supervised learning approach to solve the NLD problem and shows that we are really dealing with a *discovery* and not a learning problem.

There is a significant body of work in data mining that deals with measuring the interestingness of discovered association or classification rules [4, 8, 9]; however, these interestingness measures are not appropriate for the NLD problem. The reasons are twofold. First, most of these methods assume the data is in the form of a feature-vector (a single relational table), while for the NLD problem we have to be able to handle multi-relational data with potentially large vocabularies of relations. The second and more serious problem is that one has to first learn a pattern or rule before its interestingness can be measured. This, however, is only possible if there are enough supporting cases in the data to warrant the discovery of a particular rule. If a pattern of interest occurs only once, no rule or pattern would be available to be evaluated with one of these measures. These measures are therefore not directly applicable for novel link discovery.

3. Novel link discovery via rarity analysis

In this section we propose a set of *rarity* measures to capture the notion of "interestingness". These measures

form the foundation on which all our novel path, novel loop and significant node discovery algorithms are based.

3.1 Novel path discovery

Besides the challenges described in the previous section, another problem for novel path discovery is that the interestingness of a path is non-linearly related to the interestingness of its individual links. That is, each individual link of a path might not be interesting at all but it is the combination of them that represents something special. This non-linearity characteristic limits the effectiveness of a shortest-path like algorithm that might simply accumulate statically assigned link interestingness to compute path interestingness.

To deal with novel path discovery problems, we observe that to some extent **rarity** carries the information of **interestingness.** That is, an event that occurs infrequently compared to other events has the potential to be interesting and, thus, worth being reported. Using rarity as a measure for interestingness fulfills the need of capturing domain specificity: the same event can be rare in one domain but not in the other. For example, the event "A cites B's paper" could be interesting in a criminal database because it occurs rarely, despite the fact that people might think it to be uninteresting, since in general this citation behavior is not rare. Rarity is also flexible enough to handle different points of view. For example "A cites B's paper" can be rare from A's point of view but not from B's point of view due to the fact that A rarely cites others but B is commonly cited by many other people.

To apply these ideas to the novel path discovery problem, we have to define rarity measurements for paths in the network. Note that in a multi-relational network as shown in Figure 1, every path occurs exactly once, thus all of them are equally rare. We therefore need a more relaxed definition to measure path rarity. We do this by defining the rarity of a path as the reciprocal of the number of **similar** paths to it. We accommodate view dependency by defining four different measures based on different views of similarity.

An *n*-step path can in general be defined as a combination of *n*+1 entities (or nodes) e_i and *n* relations (or links) r_i between them:

$$e_0 \xrightarrow{r_0} e_1 \xrightarrow{r_1} e_2 \ldots\ldots \xrightarrow{r_{n-1}} e_n$$

Note that in the novel path discovery problem we do not consider paths that contain loops (in other words all *n* entities in a path must be distinct).

We define the **type** of a path as the ordered sequences of relations $[r_0 \ldots r_{n-1}]$ of that path. For example, the path "A writes a paper that cites a paper published at time t1" and the path "B writes a paper that cites a paper published at time t2" are of the same type [writes, cites, published_at].

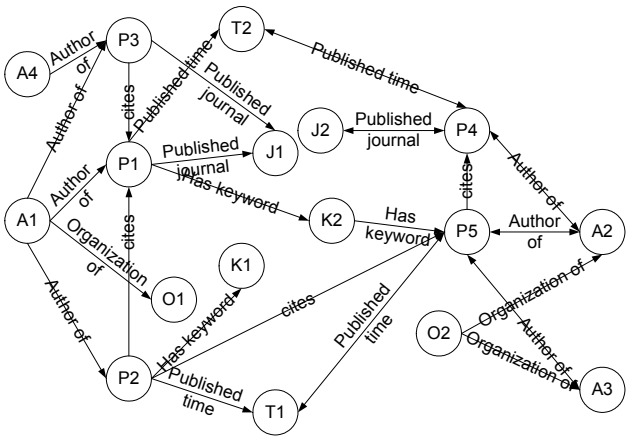


Figure 1: Example bibliography dataset containing 16 nodes and 21 links

The first path rarity measurement considers two paths as similar if they have the same type as well as identical source and target nodes. Then the rarity of a path *P* can be defined as 1/N1, where N1 is the total number of paths in the dataset that are similar to *P* in this sense. For example, in Figure 1 the path p_1 "A1 is the author of P2 and P2 cites P1" (between A1 and P1) has rarity 1/2 since there exists only one other similar path "A1 is the author of P3 and P3 cites P1". For convenience we call N1 "spindle fan-out value" since according to the constraints the path emanates from the source and terminates to the target just like a spindle.

The second path rarity measurement considers two paths as similar if they have the same type and emanate from the same source node. The rarity of a path *P* can then be defined as 1/N2, where N2 is the total number paths in the dataset that are similar to *P* in this sense. According to this rarity measure, the path p_1 described above has rarity 1/3, since there is one more path "A1 is the author of P2 and P2 cites P5" that matches the similarity criteria. We call N2 the "source fan-out value", since similar paths fans out from the source.

The third measure 1/N3 is similar to the previous one but with identical target instead of source. The rarity of path p_1 in this sense is 1/3, since besides the paths that satisfy N1 rarity, there is one more path "A4 is the author of P3 and P3 cites P1" that matches the criteria. N3 is called "target fan-out value", since similar paths fan out from the same target.

The fourth path rarity measure considers two paths with the same type as being similar. This rarity measure is defined as 1/N4 where N4 equals the total number of paths of the same type in the dataset. According to this measure, the rarity of p_1 is 1/5 since there are five paths in Figure 1 of the type "*X* is the author of *Y* and *Y* cites *Z*". We call N4 the "global fan-out value", since it represents how rare this type of path is in general.

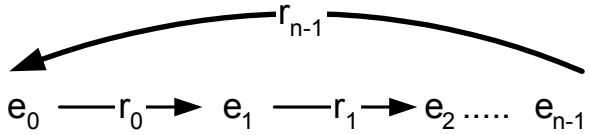
Equipped with these measures, we can now answer novel path discovery questions for the graph displayed in

Figure 1. For example: Is the path "A1 writes P2 and P2 cites P1" more interesting than the path "A1 writes P3, P3 is published in journal J1 and J1 also contains P1"? This query will have different answers for different points of view (spindle, source, target, global fan out), and which view is chosen will depend on the user's focus. For example, if the user is the author of P1, he/she might be interested in viewing things from P1 and using 1/N3 as the rarity measure. Therefore, he/she could discover that the first path is more interesting than the second given that not a lot of people in the dataset cite the paper, but many of them have papers published in the same journal. In general, 1/N2 can be used when the user cares more about the source than the target and 1/N3 is used vice versa. 1/N1 is used when the user focuses on both the source and the target. 1/N4 is used when the user is concerned more about the general rarity of the path-type (or pattern) without focusing on any individual nodes. Sometimes the query itself determines the view as well. For example, it is reasonable to use the target fan-out value when being asked whether "K1 is the keyword of paper P1 and P1's author belongs Organization O1" is more interesting than "A1 belongs to O1", since both paths ends at the same node.

With these rarity measures, we have a systematic way to answer a query such as "what is the most interesting path between nodes *X* and *Y*?" We simply enumerate all paths between *X* and *Y* and return the one with the highest rarity value. By using rarity to determine the most interesting path, we not only take the domain specificity and user views into consideration, but also avoid being misled by the apparent meaning of the links.

3.2 Novel loop discovery

The novel loop discovery problem aims at finding interesting loops in the dataset. It is a variation of novel path discovery, since a loop can be treated as a special type of a path that has identical source and target. The rarity of a loop such as this one going from e_0 to e_0



can be measured similarly to path rarity. But since in a loop the source node is identical to the target, the N1, N2, N3 value will be the same. Thus, there are only two different loop rarity measurements: 1/N1 measures how rare a specific loop is to the source and 1/N4 determines how rare this type of loop is in general.

3.3 Significant node discovery

The significant node discovery problem aims at finding the entities most significantly connected to a given node. Our intuition is that whether two nodes are significantly connected or not depends not only on the quantity but also on the quality of paths that connect them. In other words, two nodes are significantly connected with each other if there are many interesting or rare paths between them. We therefore claim that the significance between two nodes can be measured by aggregating the rarity of paths between them. Equation 1 shows how we compute the significance of connection between two nodes A and B by accumulating the path rarity of all paths connecting them.

$$node\_significance(A,B) = \sum_{\substack{P_i \in \text{paths} \\ \text{between (A,B)}}} path\_rarity(P_i)$$

Equation 1: Computing the connection significance value between two nodes A and B.

Again, which path rarity measure needs to be applied in Equation 1 depends on different points of view. Equation 2 shows how we determine the node that is the most significantly connected with node A. Note that in this case the source fan-out value N2 is used for path rarity since we have to adopt A's point of view to judge the rarity.

$$\underset{X}{\text{argmax}}\left(\sum_{\substack{P_i \in \text{paths} \\ \text{between } (A,X)}} path\_rarity(P_i)\right) = \underset{X}{\text{argmax}}\left(\sum_{\substack{P_i \in \text{paths} \\ \text{between } (A,X)}} \frac{1}{N2(P_i)}\right)$$

Equation 2: Determining the node that is most significantly connected to a given node A.

For a specific **type** of path, the N1 value represents the total number of times it occurs between source and target. The N2 value stands for the total number of times the path occurs between the source and somebody else (this is the source fan-out value). Since 1/N2 stands for how rare a path is from the source's point of view and N1 stands for how many times the path occurs between source and target, it is easy to show that the node significance value defined in Equation 1 is equivalent to the accumulation of N1/N2 for all different types of paths. Therefore, we call the N1/N2 value of a particular path type its **contribution** to the overall significance value.

According to our definition, finding an entity that is significantly connected with the source entity A is not equivalent to finding an entity that is tightly connected with A. For example, entities A and B might have much more connections between each other than entities A and C, but entity C can still be more significant to A given that there are more rare paths between A and C.

4. Experiments

Below we describe a set of experiments aimed to illustrate the validity and usefulness of our approach. The experiments are performed on the "High Energy Physics - Theory" bibliographic database (or HEP-Th), which is a

natural dataset that was used as the experimental dataset for the KDD Cup 2003[1].

The HEP-Th dataset contains a total of 29016 papers with 1.7Gbytes of associated data. Each paper in the dataset is described by a unique ID, its authors, their e-mail addresses, paper title, the journal it appeared in, publication date, abstract and a set of other papers cited by it. The source text of each paper is also available which we ignore.

To model the data we used five different types of nodes and ten different types of links. Nodes represent paper IDs (29016), author names (12755), journal names (267), organization names (753) and publication times encoded as year/season pairs (60). Numbers in parentheses indicate the number of different entities for each type in the dataset. Organizations are not given directly but inferred from author's e-mail addresses. Different spellings of author names were not consolidated and resulted in multiple nodes.

We defined the following types of links to connect nodes:

author_of(a, p) : connects author a to his/her paper p.
date_published(p, d) : connects paper p to its publication date d.
affiliation(a, o) : connects person p to an organization o he/she belongs to.
published_in(p, j) : connects paper p to journal j it is in.
cites(p, r): connects paper p to a paper r it cites.

All of these links are viewed to be directional with an implicit inverse link, thus there are a total of 5*2 link types.

In sum there are 42871 different nodes and 461932 links in the graph representing the data. We then applied our rarity measures to identify interesting paths, loops and significant nodes in this graph.

4.1 Significant node discovery

In our first experiment we attempted to evaluate our significant node discovery method. That is, given some source node S we wanted to find other nodes of various types that were significantly connected to S. Since the nodes represent real-world entities such as people, we can then manually "verify" the computed results by investigating whether they reflected real-world, significant connections visible on the World-Wide Web. For the experiment we picked C.N. Pope as the source node S, since in this dataset he is the one with the highest number of publications (130 in total), which provides us with a rich number of connections through this node.

Table 1 lists the top three interesting nodes connected to C.N. Pope for various different node types with their significance scores relative to Pope.

[1]http://www.cs.cornell.edu/projects/kddcup/datasets.html

Table 1: Nodes significantly connected to C.N. Pope

Node Type	Top-Three Scoring Nodes (sum of path rarity)		
Person	**H. Lu (4.1)**	**M. Cvetic (2.60)**	**K.S. Stelle (0.98)**
Organization	**UTexas (3.42)**	**UMich (1.80)**	**UPenn (1.18)**
Journal	**Nucl.Phys (1.33)**	**Phys.Lett (0.30)**	**Phys.Rev (0.27)**
Time	**Spring 2000 (0.40)**	**Summer 2002 (0.37)**	**Winter1995 (0.37)**

The results show that among the 12755 people in this dataset, the one that is the most significantly connected to Pope is H. Lu, while M. Cvetic is the second and K.S. Stelle is the third. To get some further insight why these people were picked as the most significant ones, we can look at what path types contributed the most to the overall significance value. The most significant path for person entities connected to C.N. Pope is that of co-authorship. This type of path emanates from Pope a total of 332 times and ends up at H. Lu 117 times, i.e., Lu contributes 35% of them while the runners up Cvetic and Stelle contribute 42 times (12.6%) and 21 times (6.3%), respectively. The second-most significant path represents a chain of co-authorship (i.e., Person1 writes with Person2 and Person2 writes with Person3 on different papers). This path is not really rare from Pope's point of view (it occurs 34473 times). However, Cvetic was involved in it 5376 times, thus, for her this type of relation still contributes 15.6% to the overall score. It shows that a significant path is not necessarily a rare path; it could be a non-rare one but occurs frequently for a specific target. The third-most significant path represents a citation relationship. Pope cites Lu's papers much more often than those of others. Looking for organizations that are interestingly connected with Mr. Pope, we found that U. Texas A&M is the most important surpassing the second U. Michigan and third U. Pennsylvania significantly.

Next we tried to verify whether the discovered relationships actually represent important real-world relationships visible through other means. After investigating through the World-Wide Web, we found that Dr. Pope is a professor at U. Texas A&M and he was Dr. Lu's thesis advisor (1988-1994) and that Dr. Lu is currently a post-doc at U. Michigan. Dr. Cvetic is a professor at U. Pennsylvania, has similar research interests to Pope and works closely with him. Dr. Stelle is a professor of Imperial College London who has ties with Pope not only academically but also personally. For example, Dr. Pope's homepage shows a picture showing him, Dr. Stelle, and others traveling together in Afghanistan.

While this "verification" is anecdotal, it does indicate that our unsupervised method, which did not know any semantics of the entities and links in this domain, is capable of returning significant relationships that are relevant in the real world.

The rest of Table 1 describes journals and time periods significantly connected to Pope. The results show that the journal Nucl.Phys. has the highest score followed by Phys.Lett. and Phys.Rev. We checked the three types of paths that contribute the most to each of these rarity values. The most important relationship discovered and taken into account by our program is frequency of publication, which intuitively makes sense. Pope published a total of 110 journal papers and 52 of them are in Nucl.Phys. He did not publish that many papers in Phys.Lett., but a significant portion of his colleagues' papers are published there. For his connection with Phys.Rev. the program discovered that the papers cited by Mr. Pope's papers are also frequently cited by papers published in Phys.Rev. As to the time periods, Spring 2000 followed by Summer 2002 and Winter 1995 connect significantly to Mr. Pope, because various types of paths such as, for example, the publication time for his papers and the publication time for his colleagues' papers, contribute relatively highly from these nodes to Pope.

4.2 Novel path discovery

We also experimented with novel path discovery questions such as, for example, which path is the most interesting (or rarest) between two people. To determine rare paths between two known nodes, we applied 1/N1 as our rarity measure where N1 is the spindle fan-out described in Section 3.1. Looking at all paths between Pope and Lu we find the path "Pope belongs to organization *O* that has another member *P* who writes a paper together with Lu" to be the rarest according to this measure. This indicates that not many of Pope's colleagues at his university write papers with Lu, which is consistent with Lu's role as Pope's student. However, this type of path is not the rarest between Pope and Cvetic, instead "Pope co-authors a paper with Cvetic" is rarer, since Cvetic seems to write more with Pope's colleagues than with him. The examples show that our novel path discovery method can take point-of-view into account, since the computed interestingness of paths changes when the view shifts (e.g. from Lu to Cvetic in this case).

In this domain rarity of individual paths does not convey such strong semantic relationships as node significance and is harder to evaluate. In this sense the relationship between path rarity and node significance resembles the relationship between a probability density function and its corresponding probability distribution function, since the integrated probability usually carries more real-world meaning than the density function itself.

4.3 Novel loop discovery

For experiments on novel loop discovery, we calculated loop rarity via 1/N4′ where N4′ is a variation of global fan-out (see Section 3.1) with the additional requirement that source and target have to be the same node. Said differently, for each possible path type leading from a node to itself we count how often that path occurs in the dataset. The rarest, least frequent loops are listed in the top portion of Table 2, the most common loops are listed at the bottom.

Table 2: The rarest and the most common loops

Top 6 loops with highest rarity value
PaperX cites PaperX
PaperX cites Paper1→Paper1 cites PaperX
PaperX cites Paper1→Paper1 cites Paper2 → Paper2 cites PaperX
PaperX cites Paper1→Paper1 cites Paper2 → Paper2 cites Paper3 →Paper3 cites PaperX
PaperX cites (or cited by) Paper1 → Paper 1 published at Time1→ At Time1, PaperX also published.
PaperX is written by Person1 → Person 1 has another Paper1→ Paper1 published at the same time as PaperX
Bottom 3 loops with lowest rarity value
PaperX cites PaperY → PaperY is being cited by PaperZ → PaperZ is being cited by PaperX
PaperX cites PaperY → PaperY published in the same journal as PaperZ→ PaperZ cites PaperX
PaperX is cited by Paper Y → Paper Y published in same season as PaperZ → PaperZ cites Paper X

The rarest loops are papers citing themselves directly, which only occurs 28 times in the whole dataset. We do not have a real world explanation for this and can only attribute it to errors in the dataset. The second, third and fourth loops are also citation loops of different length. The explanation behind this finding is that for a paper to cite another, the cited paper needs to be published earlier. In this sense a citation loop such as "P1 cites P2 cites P3 cites P1" is really a temporal contradiction and should not occur at all. One explanation for such "contradictions" is that sometimes an author (or close colleague) might cite one of his/her own submitted but not yet published papers P2 (which has already cited P1) in a paper P1. The other explanation is that one journal might have a very long revising period and during that period other people can access the previous version. For both explanations we have found supporting instances from the dataset. The fifth path shows a similar concept where it is rare for a paper to cite another paper that was published during the same time period. This type of loop could also be an indicator for authors that work closely with each other. Finally, the last path shows that people seldom publish multiple papers at the same time.

The bottom portion of the table shows the most frequently occurring loops as a contrast to the rare loops described above. For example, the most frequent loops are two papers published at the same time period that both cite X. They are loops that intuitively should occur very frequently. Note that "A cites B cites C" is a very common path, thus, we did not expect it to be interesting as a loop and were surprised by the results.

The experiments demonstrate that our approach is capable of uncovering interesting instances masked inside thousands of uninteresting facts. Furthermore, the instances found by novel loop discovery lead us to the discovery of interesting hypotheses or patterns, e.g., that citation loops are an indicator for authors who work closely with each other or for journals that have a long revision cycle.

4.4 Discussion

The experiments show that our program can find interesting connections in a network without having to learn the patterns of interestingness. For the bibliography dataset, which does not have too many different types of relations, one might be able to write a rule-based system or supervised learning program to answer similar queries as we did. However, it is time consuming to do this, since different rules or training data are required for different queries (e.g. the rules to identify the people that are interestingly connect to a keyword are different from the ones required to determine the organizations that are interestingly connected to a person). The advantage of our method is that it does everything in an unsupervised manner and eliminates the necessity to regenerate new rules or new training data for different queries or even when the whole domain is changed. It also eliminates the risk of being biased by the apparent meaning of link types.

Another advantage of our approach is that it can focus the user's attention on events that are otherwise hard to be noticed. The inspirations triggered by such evidences can sometimes lead to the discovery of pattern/knowledge. For example, without being made aware of those rare loops, we might not ever look into the issue of citation loops at all, since there are thousands of different loops in the dataset that mask this phenomenon. They also prompt us to discover other related knowledge when we try to explain them, for example, that citation loops can be an indicator of authors adding additional citations during a revision of a journal submission.

5. Related work

To our knowledge there is no other work that addresses the NLD problem in multi-relational data via an unsupervised approach. One focus of current link discovery research is on learning patterns from complex multi-relational data. For example, inductive logic programming has been applied to learn relational patterns [11] . Additionally, graph-based methods such as [6] have been used to learn subgraph categories and isomorphisms. These approaches either require training examples or learn things at the structure/schema level, while for the NLD problem it is necessary to perform discovery at the instance level by using unsupervised methods. Kovalerchuk and Vityaev's hybrid evidence correlation technique [1] first identifies common patterns via standard data mining techniques and then hypothesizes interesting or unusual patterns by negating some of the statistically significant patterns found. It is conceptually similar to our approach but requires the occurrence of very common patterns in the data.

Other analysis algorithms such as PageRank compute the importance of links through the connections between nodes in an unsupervised manner [12, 13]. In that framework, however, all relations are treated to be identical (that is, "A kills B" is not different from "A writes to B"), therefore, this approach is not suitable for the multi-relational NLD problem.

The area of outlier detection in data mining and statistics aims at detecting points that are considerably dissimilar or inconsistent with the remainder of the data [2, 3, 7, 14, 15]. This is conceptually related to our use of rarity analysis to solve the NLD problem. Current research on outlier detection, however, analyzes primarily numerical entity-attribute data instead of multi-relational social network data. In threat detection each individual event is usually not an outlier; nevertheless, combinations of seemingly harmless events can suddenly become threatening when they occur in a particular context. Outlier computations that do not take such combinations into account will fail to detect such threats. Our path rarity analysis is designed to search for these kinds of unusual connections in a multi-relational dataset.

The area of social network analysis has investigated multi-relational social behavior using graph and matrix-theoretic representations [16]. The concept of "centrality" is applied widely to determine important nodes in a network from a global point of view, while our significant node discovery tries to tackle the problem locally by answering "which node is important to a chosen node". Moreover, centrality analysis uses only the connectivity (the number of paths) to judge the significance while our algorithm considers not only the quantity but also the quality (rarity) of the paths.

Valdes-Perez [21] characterizes discovery in science as the generation of novel, interesting, plausible and intelligible knowledge about the objects of study. In this sense the novel link discovery problem is similar to literature-based discovery introduced by Swanson [18, 19], since they both intend to find interesting facts and connections in large amounts of data. Since 1986 Swanson has triggered interesting discoveries in

biomedicine strictly by looking for mediators that connect otherwise unconnected corpora of scientific literature. Literature-based discovering systems are primarily aimed at finding one-step connections between independent corpora instead of ranking the interestingness of the multi-step paths in a multi-relational network, and are therefore different from our approach.

6. Conclusion

We presented an unsupervised link discovery method aimed at detecting interesting paths or interestingly connected nodes in multi-relational datasets. Interestingness is modeled via different measures of rarity that are based on computing how often similar paths occur in the data. The method does not rely on any pre-existing or learnable pattern information and can detect novel, interesting connections that do not need to be conceived prior to the analysis. Our approach is a general-purpose method and can be applied to arbitrary multi-relational datasets. Potential applications are in law enforcement, threat detection, data cleaning [5] and scientific discovery. The experiment shows that our approach can capture interesting connections that are representative of meaningful real-world relationships. Future work will include more extensive evaluation with different data sets, handling of temporal information, negation and better handling of noise and corruption.

7. Acknowledgements

This research was supported by the Defense Advance Research Projects Agency under Air Force Research Laboratory contract F30602-01-2-0583.

8. References

[1] B. Kovalerchuk, E. Vityaev. *Correlation of complex evidences and link discovery*. *The Fifth International Conference on Forensic Statistics*. 2002. Venice, Italy.

[2] C. Aggarwal, P.Yu. *Outlier detection for high dimensional data*. *ACM SIGMOD Conference*. 2001.

[3] E.M. Knorr, R.T. Ng. *Algorithms for Mining Distance-Based Outliers in Large Datasets*. *Proc. of VLDB Conf.* 1998.

[4] A. A. Freitas, *On rule interestingness measures*. *Knowledge-Based Systems*. 1999.

[5] R.Kimball, *Dealing with Dirty Data*, *DBMS Magazine*. 1996.

[6] L. B. Holder, D. J Cook, *Graph-based data mining.* IEEE Intelligent Systems 15, 2000.

[7] M.M. Breunig, H.P. Kriegel, R. T. Ng, and J. Sander. *Optics-of: Identifying local outliers*. *Proc. of PKDD '99*.

[8] P.N. Tan, V. Kumar. *Interestingness measures for association patterns: A perspective*. *KDD*. 2000.

[9] R. Hilderman , H. Hamilton, *Knowledge discovery and interestingness measures: A survey.* 1999, Technical Report, University of Regina:

[10] R. J. Mooney, P. Melville, L. P. Rupert Tang, J. Shavlik, I.d. Dutra, D. Page, V. S. Costa. *Relational Data Mining with Inductive Logic Programming for Link Discovery*. *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*. 2002.

[11] R.J. Mooney, P.Melville, L. P. Rupert Tang, J. Shavlik, I.d. Dutra, D. Page, V. S. Costa. *Relational Data Mining with Inductive Logic Programming for Link Discovery*. *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*. 2002.

[12] S. Brin, L. Page. *The anatomy of a large-scale hypertextual Web search engine*. *Proceedings of the 7th International World Wide Web Conference*. 1998.

[13] S. Candan, W.S. Li, *Reasoning for Web Document Associations and Its Applications in Site Map Construction.* International Journal of Data and Knowledge Engineering, 2002: p.121-150.

[14] S. Ramaswamy, R. Rastogi, K. Shim. *Efficient algorithms for mining outliers from large data sets*. *Proceedings of the ACM SIGMOD Conference*. 2000.

[15] S. Shekhar, C.T. Lu, P. Zhang. *Detecting Graph-based Spatial Outliers: Algorithms and Applications*. *The Seventh ACM SIGKDD* 2001.

[16] S. Wasserman, K. Faust, *Social Network Analysis: Methods & Applications.* 1994: Cambridge, UK: Cambridge University Press.

[17] T. Senator, *Evidence Extraction and Link Discovery Program*. 2002, DARPATech 2002: http://www.darpa.mil/DARPATech2002/presentations/iao_pdf/speeches/SENATOR.pdf

[18] D. R. Swanson, *Fish Oil, Raynaud's syndrome and undiscovered public knowledge.* Perspectives in Biology and Medicine, 1986.

[19] D. R. Swanson, *Somatomedin C and arginine: Implicit connections between mutually isolated literatures.* Perspectives in Biology and Medicine, 1990.

[20] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *The KDD Process for Extracting Useful Knowledge from Volumes of Data.* Communications of the ACM, 1996. p.27-34.

[21] R. E. Valdes-Perez, *Principles of human-computer collaboration for knowledge discovery in science.* Artificial Intelligence, 1999.